

ITERATIVE ISOTONIC REGRESSION

Arnaud GUYADER¹

Université Rennes 2, INRIA and IRMAR
Campus de Villejean, Rennes, France
arnaud.guyader@uhb.fr

Nick HENGARTNER

Los Alamos National Laboratory
NM 87545, USA
nickh@lanl.gov

Nicolas JÉGOU

Université Rennes 2
Campus de Villejean, Rennes, France
nicolas.jegou@uhb.fr

Eric MATZNER-LØBER

Université Rennes 2
Campus de Villejean, Rennes, France
eml@uhb.fr

Abstract

This article introduces a new nonparametric method for estimating a univariate regression function of bounded variation. The method exploits the Jordan decomposition which states that a function of bounded variation can be decomposed as the sum of a non-decreasing function and a non-increasing function. This suggests combining the backfitting algorithm for estimating additive functions with isotonic regression for estimating monotone functions. The resulting iterative algorithm is called Iterative Isotonic Regression (I.I.R.). The main technical result in this paper is the consistency of the proposed estimator when the number of iterations k_n grows appropriately with the sample size n . The proof requires two auxiliary results that are of interest in and by themselves: firstly, we generalize the well-known consistency property of isotonic regression to the framework of a non-monotone regression function, and secondly, we relate the backfitting algorithm to Von Neumann's algorithm in convex analysis.

Index Terms — Nonparametric statistics, isotonic regression, additive models, metric projection onto convex cones.

2010 Mathematics Subject Classification: 52A05, 62G08, 62G20.

¹Corresponding author.

1 Introduction

Consider the regression model

$$Y = r(X) + \varepsilon \quad (1)$$

where X and Y are real-valued random variables, with X distributed according to a non-atomic law μ on $[0, 1]$, $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[\varepsilon|X] = 0$. We want to estimate the regression function r , assuming it is of bounded variation. Since μ is non-atomic, we will further assume, without loss of generality, that r is right-continuous. The Jordan decomposition states that r can be written as the sum of a non-decreasing function u and a non-increasing function b

$$r(x) = u(x) + b(x). \quad (2)$$

The underlying idea of the estimator that we introduce in this paper consists in viewing this decomposition as an additive model involving the increasing and the decreasing parts of r . This leads us to propose an “Iterative Isotonic Regression” estimator (abbreviated to I.I.R.) that combines the isotonic regression and backfitting algorithms, two well-established algorithms for estimating monotone functions and additive models, respectively.

The Jordan decomposition (2) is not unique in general. However, if one requires that both terms on the right-hand side have singular associated Stieltjes measures and that

$$\int_{[0,1]} r(x)\mu(dx) = \int_{[0,1]} u(x)\mu(dx), \quad (3)$$

then the decomposition is unique and the model is identifiable. Let us emphasize that, from a statistical point of view, our assumption on r is mild. The classical counterexample of a function that is not of bounded variation is $r(x) = \sin(1/x)$ for $x \in (0, 1]$, with $r(0) = 0$.

Estimating a monotone regression function is the archetypical shape restriction estimation problem. Specifically, assume that the regression function r in (1) is non-decreasing, and suppose we are given a sample $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of i.i.d. $\mathbb{R} \times \mathbb{R}$ valued random variables distributed as a generic pair (X, Y) . Then denote $x_1 = X_{(1)} < \dots < x_n = X_{(n)}$, the ordered sample and y_1, \dots, y_n the corresponding observations. In this framework, the Pool-Adjacent-Violators Algorithm (PAVA) determines a collection of non-decreasing level sets solution to the least square minimization problem

$$\min_{u_1 \leq \dots \leq u_n} \frac{1}{n} \sum_{i=1}^n (y_i - u_i)^2. \quad (4)$$

These estimators have raised great interest in the literature for decades since they are non-parametric, data driven and easy to implement. Early work on the maximum likelihood estimators of distribution parameters subject to order restriction date back to the 50's,

starting with Ayer *et al.* [2] and Brunk [6]. Comprehensive treatises on isotonic regression include Barlow *et al.* [3] and Robertson *et al.* [28]. For improvements and extensions of the PAVA approach to more general order restrictions, see Best and Chakravarti [5], Dykstra [10], and Lee [21], among others.

The solution of (4) can be seen as the metric projection, with respect to the Euclidean norm, of the vector $y = (y_1, \dots, y_n)$ on the isotone cone \mathcal{C}_n^+

$$\mathcal{C}_n^+ = \{u = (u_1, \dots, u_n) \in \mathbb{R}^n : u_1 \leq \dots \leq u_n\}. \quad (5)$$

That projection is not linear, which is the reason why analyzing these estimators is technically challenging.

Interestingly, one can interpret the isotonic regression estimator as the slope of a convex approximation of the primitive integral of r . This leads to an explicit relation between y and the vector of the adjusted values, known as the “min-max formulas” (see Anevski and Soulier [1] for a rigorous justification). This point of view plays a key role in the study of the asymptotic behavior of isotonic regression. The consistency of the estimator was established by Brunk [6] and Hanson *et al.* [15]. Brunk [7] proved its cube-root convergence at a fixed point and obtained the pointwise asymptotic distribution, and Durot [9] provided a central limit theorem for the L_p -error.

Let us now discuss the additive aspect of the model. In a multivariate setting, the additive model was originally suggested by Friedman and Stuetzle [11] and popularized by Hastie and Tibshirani [17] as a way to accommodate the so-called curse of dimensionality. The underlying idea of additive models is to approximate a high dimension regression function $r : \mathbb{R}^d \rightarrow \mathbb{R}$ by a sum of one-dimensional univariate functions, that is

$$r(\mathbf{X}) = \sum_{j=1}^d r_j(X^j). \quad (6)$$

Not only do additive models provide a logical extension of the standard linear regression model which facilitates the interpretation, but they also achieve optimal rates of convergence that do not depend on the dimension d (see Stone [29]).

Buja *et al.* [8] proposed the backfitting algorithm as a practical method for estimating additive models. It consists in iteratively fitting the partial residuals from earlier steps until convergence is achieved. Specifically, if the current estimates are $\hat{r}_1, \dots, \hat{r}_d$, then \hat{r}_j is updated by smoothing $y - \sum_{k \neq j} \hat{r}_k$ against X^j . The backfitted estimators have mainly been studied in the case of linear smoothers. Härdle and Hall [16] showed that when all the smoothers are orthogonal projections, the whole algorithm can be replaced by a global projection operator. Opsomer and Ruppert [26], and Opsomer [27], gave asymptotic bias and variance expressions in the context of additive models fitted by local

polynomial regression. Mammen, Linton and Nielsen [23] improved these results by deriving a backfitting procedure that achieves the oracle efficiency (that is, each component can be estimated as well as if the other components were known). This procedure was extended to several different one-dimensional smoothers including kernel, local polynomials and splines by Horowitz, Klemelä and Mammen [18]. Alternative estimation procedures for additive models have been considered by Kim, Linton and Hengartner [20], and by Hengartner and Sperlich [19].

In the present context, we propose to apply the backfitting algorithm to decompose a univariate function by alternating isotonic and antitonic regressions on the partial residuals in order to estimate the additive components u and b of the Jordan decomposition (2). The finite sample behavior of this estimator has been studied in a related paper by Guyader *et al.* (see [14]). Among other results, it is stated that the sequence of estimators obtained in this way converges to an interpolant of the raw data (see section 2 below for details).

Backfitted estimators in a non-linear case have also been studied by Mammen and Yu [24]. Specifically, assuming that the regression function r in (6) is an additive function of isotonic one-dimensional functions r_j , they estimate each additive component by iterating the PAVA in a backfitting fashion. Moreover, Mammen and Yu show that, as in the linear case, their estimator achieves the oracle efficiency and, in each direction, they recover the limit distribution exhibited by Brunk [7].

The main result addressed in this paper states the consistency of our I.I.R. estimator. Denoting $\hat{r}_n^{(k)}$ the Iterative Isotonic Regression estimator resulting from k iterations of the algorithm, we prove the existence of a sequence of iterations (k_n) , increasing with the sample size n , such that

$$\mathbb{E} [\|\hat{r}_n^{(k_n)} - r\|^2] \xrightarrow{n \rightarrow \infty} 0 \quad (7)$$

where $\|\cdot\|$ is the quadratic norm with respect to the law μ of X . Our analysis identifies two error terms: an estimation error that comes from the isotonic regression, and an approximation error that is governed by the number of iterations k .

Concerning the estimation error, we wish to emphasize that all asymptotic results about isotonic regression mentioned above assume monotonicity of the regression function r . In our context, at each stage of the iterative process, we apply an isotonic regression to an arbitrary function (of bounded variation). As a result, we prove in Section 3 the $L_2(\mu)$ consistency of isotonic regression for the metric projection of r onto the cone of increasing functions (see Theorem 1).

The approximation term can be controlled by increasing the number of iterations. This is made possible thanks to the interpretation of I.I.R. as a Von Neumann's algorithm, and by applying related results in convex analysis (see Proposition 3). Putting estimation and

approximation errors together finally leads to the consistency result (7).

Let us remark that, as far as we know, rates of convergence of Von Neumann's algorithm have not yet been studied in the context of bounded variation functions. Hence, at this time, it seems difficult to establish rates of convergence for our estimator without further restrictions on the shape of the underlying regression function. Thus, the results we present here may be considered as a starting point in the study of novel methods which would consist in applying isotonic regression with no particular shape assumption on the regression function.

The remainder of the paper is organised as follows. We first give further details and notations about the construction of I.I.R. in Section 2. The general consistency result for isotonic regression is given in Section 3. The main result of this article, the consistency of I.I.R., is established in Section 4. Most of the proofs are postponed to Section 5, while related technical results are gathered in Section 6.

2 The I.I.R. procedure

Denote by $y = (y_1, \dots, y_n)$ the vector of observations corresponding to the ordered sample $x_1 = X_{(1)} < \dots < X_{(n)} = x_n$. We implicitly assume in this writing that the law μ of X has no atoms. We denote by $\text{iso}(y)$ (resp. $\text{anti}(y)$) the metric projection of y with respect to the Euclidean norm onto the isotone cone \mathcal{C}_n^+ (resp. $\mathcal{C}_n^- = -\mathcal{C}_n^+$) defined in (5):

$$\begin{aligned}\text{iso}(y) &= \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - u_i)^2 = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|y - u\|_n^2 \\ \text{anti}(y) &= \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - b_i)^2 = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|y - b\|_n^2.\end{aligned}$$

The backfitting algorithm consists in updating each component by smoothing the partial residuals, i.e., the residuals resulting from the current estimate in the other direction. Thus the Iterative Isotonic Regression algorithm goes like this:

Algorithm 1 Iterative Isotonic Regression (I.I.R.)

- (1) Initialization: $\hat{b}_n^{(0)} = (\hat{b}_1^{(0)}[1], \dots, \hat{b}_n^{(0)}[n]) = 0$
- (2) Cycle: for $k \geq 1$

$$\begin{aligned}\hat{u}_n^{(k)} &= \text{iso} \left(y - \hat{b}_n^{(k-1)} \right) \\ \hat{b}_n^{(k)} &= \text{anti} \left(y - \hat{u}_n^{(k)} \right) \\ \hat{r}_n^{(k)} &= \hat{u}_n^{(k)} + \hat{b}_n^{(k)}.\end{aligned}$$

- (3) Iterate (2) until a stopping condition to be specified is achieved.
-

Guyader *et al.* [14] prove that the terms of the decomposition $\hat{r}_n^{(k)} = \hat{u}_n^{(k)} + \hat{b}_n^{(k)}$ have singular Stieltjes measures. Furthermore, by starting with isotonic regression, the terms $\hat{u}_n^{(k)}$ have all the same empirical mean as the original data y , while all the $\hat{b}_n^{(k)}$ are centered. Hence, for each k , the decomposition $\hat{r}_n^{(k)} = \hat{u}_n^{(k)} + \hat{b}_n^{(k)}$ satisfies the condition (3), and that decomposition is unique (identifiable).

Algorithm 1 furnishes vectors of adjusted values. In the following, we will consider one-to-one mappings between such vectors and piecewise functions defined on the interval $[0, 1]$. For example, the vector $\hat{u}_n^{(k)} = (\hat{u}_n^{(k)}[1], \dots, \hat{u}_n^{(k)}[n])$ is associated to the real-valued function $\hat{u}_n^{(k)}$ defined on $[0, 1]$ by

$$\hat{u}_n^{(k)}(x) = \hat{u}_n^{(k)}[1] \mathbb{1}_{[0, X_{(2)})}(x) + \sum_{i=2}^{n-1} \hat{u}_n^{(k)}[i] \mathbb{1}_{[X_{(i)}, X_{(i+1)})}(x) + \hat{u}_n^{(k)}[n] \mathbb{1}_{[X_{(n)}, 1]}(x). \quad (8)$$

Observe that our definition of $\hat{u}_n^{(k)}(x)$ makes it right-continuous. Obviously, equivalent formulations hold for $\hat{b}_n^{(k)}$ and $\hat{r}_n^{(k)}$ as well.

Figure 1 illustrates the application of l.l.R. on an example. The top left-hand side displays the regression function r , and $n = 100$ points (x_i, y_i) , with $y_i = r(x_i) + \varepsilon_i$, where the ε_i 's are Gaussian centered random variables. The three other figures show the estimations $\hat{r}_n^{(k)}$ obtained on this sample for $k = 1, 10$, and $1,000$ iterations. According to (8), our method fits a piecewise constant function. Moreover, increasing the number of iterations tends to increase the number of jumps.

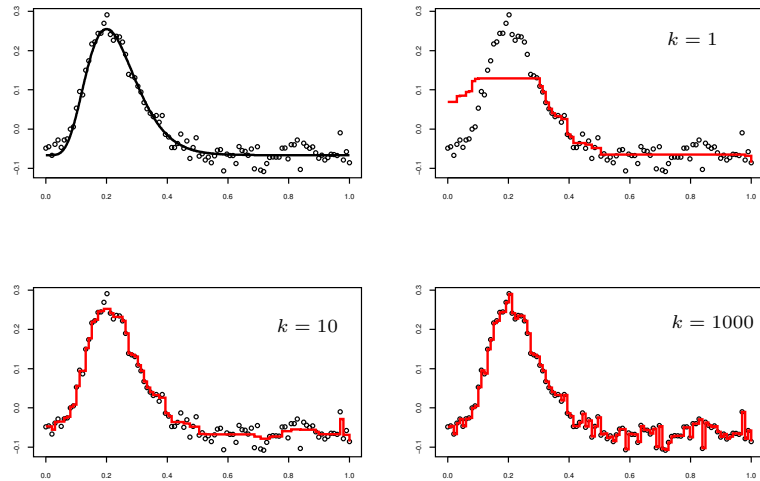


Figure 1: Application of the l.l.R. algorithm for $k = 1, 10$, and $1,000$ iterations.

The bottom right figure illustrates that, as established in Guyader *et al.* [14], for fixed sample size n , the function $\hat{r}_n^{(k)}(x)$ converges to an interpolant of the data when the

number of iterations k tends to infinity, *i.e.*, for all $i = 1, \dots, n$,

$$\lim_{k \rightarrow \infty} \hat{r}_n^{(k)}(x_i) = y_i.$$

One interpretation of the above result is that increasing the number of iterations leads to overfitting. Thus, iterating the procedure until convergence is not desirable. On the other hand, as illustrated on figure 1, iterations beyond the first step typically improve the fit. This suggests that we need to couple the l.l.R. algorithm with a stopping rule. In this respect, two important remarks are in order. Firstly, since equation (8) enables predictions at arbitrary locations $x \in [0, 1]$, all the standard data-splitting techniques can be applied to stop the algorithm.

Secondly, the choice of a stopping criterion as a model selection suggests stopping rules based on Akaike Information Criterion, Bayesian Information Criterion or Generalized Cross Validation. These criteria can be written in the generic form

$$\operatorname{argmin}_p \left\{ \log \frac{1}{n} \text{RSS}(p) + \phi(p) \right\}. \quad (9)$$

Here, RSS denotes the residual sum of squares and ϕ is an increasing function. The parameter p stands for the number (or equivalent number) of parameters. For isotonic regression, we refer to Meyer and Woodroffe [25] to consider that the number of jumps provides the effective dimension of the model. Therefore, a natural extension for l.l.R. is to replace p by the number of jumps of $\hat{r}_n^{(k)}$ in (9). The comparisons of these criteria and the practical behavior of the l.l.R. procedure will be addressed elsewhere by the authors.

3 Isotonic regression: a general result of consistency

In this section, we focus on the first half step of the algorithm, which consists in applying isotonic regression to the original data. To simplify the notations, we omit in this section the exponent related to the number of iterations k , and simply denote \hat{u}_n the isotonic regression on the data, that is,

$$\hat{u}_n = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \|y - u\|_n = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \frac{1}{n} \sum_{i=1}^n (y_i - u_i)^2.$$

Let u_+ denote the closest non-decreasing function to the regression function r with respect to the $L_2(\mu)$ norm. Thus, u_+ is defined as

$$u_+ = \operatorname{argmin}_{u \in \mathcal{C}^+} \|r - u\| = \operatorname{argmin}_{u \in \mathcal{C}^+} \int_{[0,1]} (r(x) - u(x))^2 \mu(dx),$$

where \mathcal{C}^+ denotes the cone of non-decreasing functions in $L_2(\mu)$. Since \mathcal{C}^+ is closed and convex, the metric projection u_+ exists and is unique in $L_2(\mu)$.

For mathematical purpose, we also introduce u_n , the result from applying isotonic regression to the sample $(x_i, r(x_i))$, $i = 1, \dots, n$, that is

$$u_n = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \|r - u\|_n = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \frac{1}{n} \sum_{i=1}^n (r(x_i) - u_i)^2. \quad (10)$$

Finally, we note that, since r is bounded, so are u_+ and u_n , independently of the sample size n (see for example Lemma 2 in Anevski and Soulier [1]). Figure 2 displays the three terms involved.

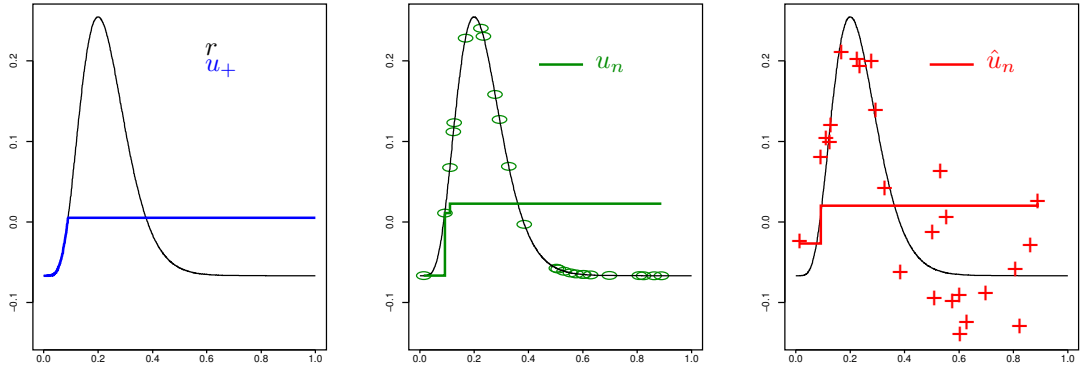


Figure 2: Isotonic regression on a non-monotone regression function.

The main result of this section states that

$$\mathbb{E} [\|\hat{u}_n - u_+\|^2] \xrightarrow{n \rightarrow \infty} 0,$$

where the expectation is taken with respect to the sample \mathcal{D}_n . Our analysis decomposes $\|\hat{u}_n - u_+\|$ into two distinct terms:

$$\|\hat{u}_n - u_+\| \leq \|\hat{u}_n - u_n\| + \|u_n - u_+\|.$$

As $\|u_n - u_+\|$ does not depend on the response variable Y_i , one could interpret it as a bias term, whereas $\|\hat{u}_n - u_n\|$ plays the role of a variance term.

Throughout this section, our results are stated for both the empirical norm $\|\cdot\|_n$ and the $L_2(\mu)$ norm $\|\cdot\|$, as both are informative. The following proposition states the convergence of the bias term (its proof is postponed to Section 5.1).

Proposition 1 *With the previous notations, we have*

$$\lim_{n \rightarrow \infty} \|u_n - u_+\|_n = 0 \quad a.s.,$$

and

$$\lim_{n \rightarrow \infty} \|u_n - u_+\| = 0 \quad a.s.$$

Applying Lebesgue's dominated convergence Theorem ensures that both

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|u_n - u_+\|_n^2] = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{E} [\|u_n - u_+\|^2] = 0.$$

Analysis of the variance term requires that we assume that the noise ε is bounded. It then follows from Anevski and Soulier [1] that \hat{u}_n is bounded, independently of the sample size n . The proof of the following result is given in Section 5.2).

Proposition 2 *Assume that the random variable ε is bounded, then we have*

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n - u_n\|_n^2] = 0,$$

and

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n - u_n\|^2] = 0.$$

Combining Proposition 1 and Proposition 2 yields the following theorem.

Theorem 1 *Consider the model $Y = r(X) + \varepsilon$, where $r : [0, 1] \rightarrow \mathbb{R}$ belongs to $L_2(\mu)$, μ is a non-atomic distribution on $[0, 1]$, and ε is a bounded random variable satisfying $\mathbb{E}[\varepsilon|X] = 0$. Denote u_+ and \hat{u}_n the functions resulting from the isotonic regression applied on r and on the sample \mathcal{D}_n , respectively. Then we have*

$$\mathbb{E} [\|\hat{u}_n - u_+\|_n^2] \rightarrow 0$$

and

$$\mathbb{E} [\|\hat{u}_n - u_+\|^2] \rightarrow 0$$

when the sample size n tends to infinity.

This result generalizes the consistency of isotonic regression when applied in a more general context than the one of monotone functions. It will be of constant use when iterating our algorithm. This is the topic of the upcoming section.

4 Consistency of iterative isotonic regression

We now proceed with our main result, which states that there is a sequence of iterations k_n , increasing with the sample size n , such that

$$\mathbb{E} [\|\hat{r}_n^{(k_n)} - r\|^2] \xrightarrow{n \rightarrow \infty} 0.$$

In order to control the expectation of the L_2 distance between the estimator $\hat{r}_n^{(k)}$ and the regression function r , we shall split $\|\hat{r}_n^{(k)} - r\|$ as follows: let $r^{(k)}$ be the result from applying the algorithm on the regression function r itself k times, that is $r^{(k)} = u^{(k)} + b^{(k)}$, where

$$u^{(k)} = \operatorname{argmin}_{u \in \mathcal{C}^+} \|r - b^{(k-1)} - u\| \quad \text{and} \quad b^{(k)} = \operatorname{argmin}_{b \in \mathcal{C}^-} \|r - u^{(k)} - b\|.$$

We then upper-bound

$$\|\hat{r}_n^{(k)} - r\| \leq \|r^{(k)} - r\| + \|\hat{r}_n^{(k)} - r^{(k)}\|. \quad (11)$$

In this decomposition, the first term is an approximation error, while the second one corresponds to an estimation error.

Figure 3 displays the function $r^{(k)}$ for two particular values of k . One can see that, after k steps of the algorithm, there generally remains an approximation error $\|r^{(k)} - r\|$. Nonetheless, one also observes that this error decreases when iterating the algorithm.

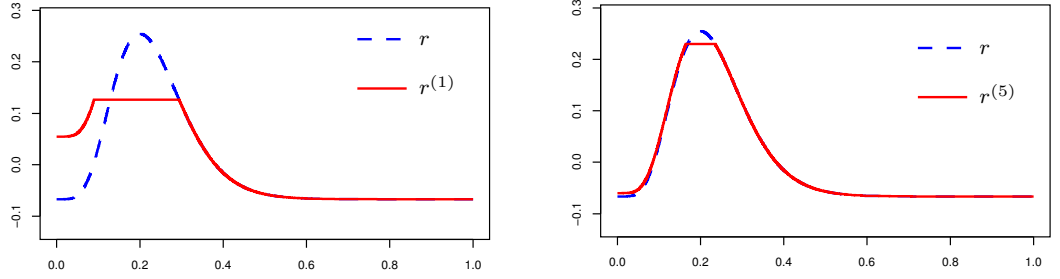


Figure 3: Decreasing of the approximation error $\|r^{(k)} - r\|$ with k .

The following proposition states that the approximation error can indeed be controlled by increasing the number of iterations k . Its proof relies on the interpretation of I.I.R. as a Von Neumann's algorithm (see Section 5.3 for the proof).

Proposition 3 *Assume that r is a right-continuous function of bounded variation and μ a non-atomic law on $[0, 1]$. Then the approximation term $\|r^{(k)} - r\|$ tends to 0 when the number of iterations grows:*

$$\lim_{k \rightarrow \infty} \|r^{(k)} - r\| = 0,$$

where $\|\cdot\|$ denotes the quadratic norm in $L_2(\mu)$.

Coming back to (11), we further decompose the estimation error into a bias and a variance term to obtain

$$\begin{aligned} \|\hat{r}_n^{(k)} - r\| &\leq \underbrace{\|\hat{r}_n^{(k)} - r^{(k)}\|}_{\text{Estimation}} + \underbrace{\|r^{(k)} - r\|}_{\text{Approximation}} \\ &\leq \underbrace{\|\hat{r}_n^{(k)} - r_n^{(k)}\|}_{\text{Variance}} + \underbrace{\|r_n^{(k)} - r^{(k)}\|}_{\text{Bias}} \end{aligned}$$

The function $r_n^{(k)}$ results from k iterations of the algorithm on the sample $(x_i, r(x_i))$, $i = 1, \dots, n$, and can be seen as the equivalent of the function u_n defined in (10). This decomposition allows us to make use of the consistency results of the previous section, and to control the estimation error when the sample size n goes to infinity. We now state the main theorem of this paper.

Theorem 2 *Consider the model $Y = r(X) + \varepsilon$, where $r : [0, 1] \rightarrow \mathbb{R}$ is a right-continuous function of bounded variation, μ a non-atomic distribution on $[0, 1]$, and ε a bounded random variable satisfying $\mathbb{E}[\varepsilon|X] = 0$. Then there exists an increasing sequence of iterations (k_n) such that*

$$\mathbb{E} [\|\hat{r}_n^{(k_n)} - r\|^2] \xrightarrow{n \rightarrow \infty} 0,$$

where $\|\cdot\|$ denotes the quadratic norm in $L_2(\mu)$.

Proof. Coming back to the original notation, Theorem 1 states that

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n^{(1)} - u^{(1)}\|_n^2] = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n^{(1)} - u^{(1)}\|^2] = 0. \quad (12)$$

In the following, we show that this result still holds when applying the backfitting algorithm. Before proceeding, just remark that, since r and ε are bounded, this will also be the case for all the quantities at stake in the remainder of the proof. In particular, this allows us to use the concentration inequalities established in Section 6.1.

- We first describe the end of the first step by showing that $\mathbb{E} [\|\hat{b}_n^{(1)} - b^{(1)}\|^2] \rightarrow 0$.

Recall the definitions

$$b^{(1)} = \operatorname{argmin}_{b \in \mathcal{C}^-} \|r - u^{(1)} - b\| \quad \text{and} \quad \hat{b}_n^{(1)} = \operatorname{argmin}_{b \in \mathcal{C}_n^-} \|y - \hat{u}_n^{(1)} - b\|_n.$$

In order to mimic the previous step, let us consider the vectors

$$\tilde{y} = y - u^{(1)} \quad \text{and} \quad \tilde{b}_n^{(1)} = \operatorname{argmin}_{b \in \mathcal{C}_n^-} \|\tilde{y} - b\|_n,$$

so that

$$\tilde{y} = (r - u^{(1)}) + \varepsilon$$

and

$$\tilde{b}_n^{(1)} = \operatorname{argmin}_{b \in \mathcal{C}_n^-} \|(r - u^{(1)}) + \varepsilon - b\|_n.$$

To study the term $\|\tilde{b}_n^{(1)} - b^{(1)}\|$, one can apply *mutatis mutandis* the result of Theorem 1, replacing $\hat{u}_n^{(1)}$ by $\tilde{b}_n^{(1)}$, r by $r - u^{(1)}$, and isotonic regression by antitonic regression. Hence,

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\tilde{b}_n^{(1)} - b^{(1)}\|_n^2] = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{E} [\|\tilde{b}_n^{(1)} - b^{(1)}\|^2] = 0. \quad (13)$$

As projection reduces distances, we also have

$$\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n \leq \|y - \hat{u}_n^{(1)} - \tilde{y}\|_n = \|\hat{u}_n^{(1)} - u^{(1)}\|_n.$$

Thanks to equations (12) and (13), we deduce

$$\mathbb{E} [\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2] \leq 2 \times \left\{ \mathbb{E} [\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n^2] + \mathbb{E} [\|\tilde{b}_n^{(1)} - b^{(1)}\|_n^2] \right\} \rightarrow 0.$$

Invoking the same arguments as those at the end of the proof of Proposition 2, we also have

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{b}_n^{(1)} - b^{(1)}\|^2] = 0.$$

Finally, at the end of the first iteration, we have

$$\mathbb{E} [\|\hat{r}_n^{(1)} - r^{(1)}\|^2] \leq 2 \times \left\{ \mathbb{E} [\|\hat{u}_n^{(1)} - u^{(1)}\|^2] + \mathbb{E} [\|\hat{b}_n^{(1)} - b^{(1)}\|^2] \right\} \rightarrow 0.$$

• For the beginning of the second iteration, consider this time

$$\hat{u}_n^{(2)} = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \|y - \hat{b}_n^{(1)} - u\|_n \quad \text{and} \quad u^{(2)} = \operatorname{argmin}_{u \in \mathcal{C}^+} \|r - b^{(1)} - u\|.$$

Let us introduce

$$\tilde{y} = y - b^{(1)} = (r - b^{(1)}) + \varepsilon \quad \text{and} \quad \tilde{u}_n^{(2)} = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \|\tilde{y} - u\|_n = \operatorname{argmin}_{u \in \mathcal{C}_n^+} \|(r - b^{(1)}) + \varepsilon - u\|_n.$$

We apply Theorem 1 again, replacing r by $r - b^{(1)}$, and $\hat{u}_n^{(1)}$ by $\tilde{u}_n^{(2)}$. This leads to

$$\lim_{n \rightarrow 0} \mathbb{E} [\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2] = 0.$$

Thanks to the reduction property of isotonic regression and using the conclusion of the first iteration, we get

$$\mathbb{E} [\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2] \leq \mathbb{E} [\|y - \hat{b}_n^{(1)} - ((r - b^{(1)}) + \varepsilon)\|_n^2] = \mathbb{E} [\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2] \rightarrow 0.$$

Therefore

$$\mathbb{E} [\|\hat{u}_n^{(2)} - u^{(2)}\|_n^2] \leq 2 \times \left\{ \mathbb{E} [\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2] + \mathbb{E} [\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2] \right\} \rightarrow 0$$

and, as before, we also have

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n^{(2)} - u^{(2)}\|^2] = 0.$$

The same scheme leads to $\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{b}_n^{(2)} - b^{(2)}\|^2] = 0$, so that

$$\mathbb{E} [\|\hat{r}_n^{(2)} - r^{(2)}\|^2] \leq 2 \times \left\{ \mathbb{E} [\|\hat{u}_n^{(2)} - u^{(2)}\|^2] + \mathbb{E} [\|\hat{b}_n^{(2)} - b^{(2)}\|^2] \right\} \rightarrow 0.$$

• By iterating this process, it is readily seen that, for all $k \geq 1$,

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{r}_n^{(k)} - r^{(k)}\|^2] = 0,$$

which means that, at each iteration, the estimation error goes to 0 when the sample size tends to infinity.

We deduce that we can construct an increasing sequence (n_k) such that for each $k \geq 1$ and for all $n \geq n_k$

$$\mathbb{E} [\|\hat{r}_n^{(k)} - r^{(k)}\|] \leq \|r^{(k)} - r\| + \frac{1}{k}.$$

Notice that the term $\|r^{(k)} - r\|$ might be equal to zero (*e.g.*, $r^{(1)} = r$ if r is monotone), hence the additive term $1/k$ in the previous inequality. Consequently,

$$\mathbb{E} [\|\hat{r}_n^{(k)} - r\|] \leq 2\|r^{(k)} - r\| + \frac{1}{k}.$$

Then let us consider the sequence (k_n) defined as: $k_n = 0$ if $n < n_1$, $k_n = 1$ if $n_1 \leq n < n_2$, and so on. Obviously (k_n) tends to infinity and

$$\mathbb{E} [\|\hat{r}_n^{(k_n)} - r\|] \leq 2\|r^{(k_n)} - r\| + \frac{1}{k_n} \xrightarrow{n \rightarrow \infty} 0.$$

This ends the proof of Theorem 2. □

5 Proofs

5.1 Proof of Proposition 1

For g and h two functions defined on $[0, 1]$, we denote $\Delta_n(g - h)$ the random variable

$$\Delta_n(g - h) = \|g - h\|_n^2 - \|g - h\|^2 = \frac{1}{n} \sum_{i=1}^n \{(g(X_i) - h(X_i))^2 - \mathbb{E} [(g(X) - h(X))^2]\}.$$

We first show that

$$\|r - u_n\|_n \rightarrow \|r - u_+\| \quad a.s. \quad (14)$$

To this end, we proceed in two steps, proving in a first time that

$$\limsup \|r - u_n\|_n \leq \|r - u_+\| \quad a.s. \quad (15)$$

and in a second time that

$$\liminf \|r - u_n\|_n \geq \|r - u_+\| \quad a.s. \quad (16)$$

For the first inequality, let us denote

$$A_n = \{|\Delta_n(r - u_+)| > n^{-1/3}\} = \{|\|r - u_+\|_n^2 - \|r - u_+\|^2| > n^{-1/3}\}.$$

By the definition of u_n , note that for all n ,

$$\|r - u_n\|_n \leq \|r - u_+\|_n$$

so that on $\overline{A_n}$,

$$\|r - u_n\|_n^2 \leq \|r - u_+\|_n^2 \leq \|r - u_+\|^2 + n^{-1/3}.$$

Consequently

$$B_n = \{\|r - u_n\|_n^2 \leq \|r - u_+\|^2 + n^{-1/3}\} \supset \overline{A_n}.$$

Therefore

$$\mathbb{P}(\liminf B_n) \geq \mathbb{P}(\liminf \overline{A_n}) = 1 - \mathbb{P}(\limsup A_n).$$

Invoking Lemma 1 and Borel-Cantelli Lemma, we conclude that $\mathbb{P}(\limsup A_n) = 0$, and hence $\mathbb{P}(\liminf B_n) = 1$. On the set $\liminf B_n$, we have

$$\limsup \|r - u_n\|_n^2 \leq \|r - u_+\|^2,$$

which proves Equation (15).

Conversely, we now establish Equation (16). By definition of u_+ , observe that for all n ,

$$\|r - u_+\| \leq \|r - u_n\|.$$

Consider the sets

$$C_n = \left\{ \sup_{h \in \mathcal{C}_{[a,b]}^+} |\Delta_n(r - h)| > n^{-1/3} \right\} \text{ and } D_n = \{\|r - u_n\|_n^2 \geq \|r - u_+\|^2 - n^{-1/3}\}$$

so that $\overline{C_n} \subset D_n$, and by applying Lemma 2,

$$\mathbb{P}(\liminf D_n) \geq 1 - \mathbb{P}(\limsup C_n) = 1.$$

On the set $\liminf D_n$, one has

$$\liminf \|r - u_n\|_n^2 \geq \|r - u_+\|^2,$$

which proves (16). Combining Equations (15) and (16) leads to (14).

Next, using Lemma 2 again, we get

$$\lim_{n \rightarrow \infty} \|r - u_n\|_n - \|r - u_n\| = 0 \quad a.s.$$

Combined with (14), this leads to

$$\|r - u_n\| \rightarrow \|r - u_+\| \quad a.s. \quad (17)$$

It remains to prove the almost sure convergence of u_n to u_+ . For this, it suffices to use the parallelogram law. Indeed, noting $m_n = (u_n + u_+)/2$, we have

$$\|u_n - u_+\|^2 = 2(\|r - u_+\|^2 + \|u_n - r\|^2) - 4\|m_n - r\|^2.$$

Since both u_+ and u_n belong to the convex set \mathcal{C}^+ , so does m_n . Hence $\|r - u_+\|^2 \leq \|r - m_n\|^2$, and

$$\|u_n - u_+\|^2 \leq 2(\|u_n - r\|^2 - \|r - u_+\|^2).$$

Combining this with (17), we conclude that

$$\lim_{n \rightarrow \infty} \|u_n - u_+\| = 0 \quad a.s.$$

Finally, Lemma 2 guarantees the same result for the empirical norm, that is

$$\lim_{n \rightarrow \infty} \|u_n - u_+\|_n = 0 \quad a.s.$$

and the proof is complete.

5.2 Proof of Proposition 2

Let us denote $\langle \cdot, \cdot \rangle_n$ the inner product associated to the empirical norm $\|\cdot\|_n$. Since isotonic regression corresponds to the metric projection onto the closed convex cone \mathcal{C}_n^+ with respect to this empirical norm, the vectors \hat{u}_n et u_n are characterized by the following inequalities: for any vector $u \in \mathcal{C}_n^+$,

$$\langle y - \hat{u}_n, u - \hat{u}_n \rangle_n \leq 0 \quad (18)$$

$$\langle r - u_n, u - u_n \rangle_n \leq 0 \quad (19)$$

Setting $u = u_n$ in (18) and $u = \hat{u}_n$ in (19), we get

$$\langle y - \hat{u}_n, u_n - \hat{u}_n \rangle_n \leq 0 \quad \text{and} \quad \langle r - u_n, \hat{u}_n - u_n \rangle_n \leq 0.$$

Since $\varepsilon = y - r$, this leads to

$$\|\hat{u}_n - u_n\|_n^2 \leq \langle \varepsilon, \hat{u}_n - u_n \rangle_n. \quad (20)$$

Next, we have to use an approximation result, namely Lemma 5 in Section 6.2. The underlying idea is to exploit the fact that any non-decreasing bounded sequence can be approached by the element of a subspace H_+ at distance less than δ . Specifically, if C is an upper-bound for the absolute value of the considered non-decreasing bounded sequences, we can construct such a subspace H_+ with dimension N where $N = (8C^2)/\delta^2$. From now on, we will take $N \leq n$. Before proceeding, just notice that the boundedness assumption on the random variables ε_i allows us to find a common upper bound C for the absolute values of the components of \hat{u}_n and u_n .

Let us introduce the vectors \hat{h}_n and h_n defined by

$$\hat{h}_n = \inf_{h \in H_+} \|\hat{u}_n - h\|_n \quad \text{and} \quad h_n = \inf_{h \in H_+} \|u_n - h\|_n$$

so that

$$\|\hat{u}_n - \hat{h}_n\|_n \leq \delta \quad \text{and} \quad \|u_n - h_n\|_n \leq \delta.$$

From this, we get

$$\begin{aligned} \langle \varepsilon, \hat{u}_n - u_n \rangle_n &= \langle \varepsilon, \hat{u}_n - \hat{h}_n \rangle_n + \langle \varepsilon, \hat{h}_n - h_n \rangle_n + \langle \varepsilon, h_n - u_n \rangle_n \\ &\leq \|\hat{h}_n - h_n\|_n \left\langle \varepsilon, \frac{\hat{h}_n - h_n}{\|\hat{h}_n - h_n\|_n} \right\rangle_n + 2\delta \|\varepsilon\|_n \\ &\leq \left\{ \|\hat{h}_n - \hat{u}_n\|_n + \|\hat{u}_n - u_n\|_n + \|u_n - h_n\|_n \right\} \sup_{v \in H_+, \|v\|_n=1} \langle \varepsilon, v \rangle_n + 2\delta \|\varepsilon\|_n \\ &\leq \{\|\hat{u}_n - u_n\|_n + 2\delta\} \sup_{v \in H_+, \|v\|_n=1} \langle \varepsilon, v \rangle_n + 2\delta \|\varepsilon\|_n. \end{aligned}$$

According to (20), we deduce

$$\|\hat{u}_n - u_n\|_n^2 \leq \{\|\hat{u}_n - u_n\|_n + 2\delta\} \sup_{v \in H_+, \|v\|_n=1} \langle \varepsilon, v \rangle_n + 2\delta \|\varepsilon\|_n$$

so that

$$\|\hat{u}_n - u_n\|_n^2 \leq \{\|\hat{u}_n - u_n\|_n + 2\delta\} \|\pi_{H_+}(\varepsilon)\|_n + 2\delta \|\varepsilon\|_n,$$

where $\pi_{H_+}(\varepsilon)$ stands for the metric projection of ε onto H_+ . Put differently, we have

$$\|\hat{u}_n - u_n\|_n^2 \leq \|\hat{u}_n - u_n\|_n \times \|\pi_{H_+}(\varepsilon)\|_n + 2\delta \{\|\pi_{H_+}(\varepsilon)\|_n + \|\varepsilon\|_n\},$$

and taking the expectation on both sides leads to

$$\mathbb{E} [\|\hat{u}_n - u_n\|_n^2] \leq \mathbb{E} [\|\hat{u}_n - u_n\|_n \times \|\pi_{H_+}(\varepsilon)\|_n] + 2\delta \{\mathbb{E} [\|\pi_{H_+}(\varepsilon)\|_n] + \mathbb{E} [\|\varepsilon\|_n]\}.$$

If we denote

$$\begin{cases} x^2 &= \mathbb{E} [\|\hat{u}_n - u_n\|_n^2] \\ \alpha_n &= \sqrt{\mathbb{E} [\|\pi_{H_+}(\varepsilon)\|_n^2]} \\ \beta_n &= 2\delta \{\mathbb{E} [\|\pi_{H_+}(\varepsilon)\|_n] + \mathbb{E} [\|\varepsilon\|_n]\} \end{cases}$$

an application of Cauchy-Schwarz inequality gives

$$x^2 - \alpha_n x - \beta_n \leq 0 \Rightarrow x \leq \frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2},$$

which means that

$$\mathbb{E} [\|\hat{u}_n - u_n\|_n^2] \leq \left(\frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2} \right)^2.$$

Since the random variables ε_i are i.i.d. with mean zero and common variance σ^2 , a straightforward computation shows that

$$\mathbb{E} [\|\pi_{H_+}(\varepsilon)\|_n^2] = \frac{1}{n} \mathbb{E} [(\pi_{H_+}\varepsilon)'(\pi_{H_+}\varepsilon)] = \frac{1}{n} \mathbb{E} [\text{tr}((\pi_{H_+}\varepsilon)'(\pi_{H_+}\varepsilon))] = \frac{1}{n} \text{tr}(\mathbb{E}[\varepsilon\varepsilon'] \pi_{H_+}),$$

and since H_+ has dimension $N = (8C^2)/\delta^2$, this gives

$$\mathbb{E} [\|\pi_{H_+}(\varepsilon)\|_n^2] = \sigma^2 \frac{N}{n} \Rightarrow \alpha_n = \sigma \sqrt{\frac{N}{n}} = \frac{2\sqrt{2}C\sigma}{\delta\sqrt{n}}.$$

Set $\delta = \delta_n = n^{-\alpha}$ with $0 < \alpha < 1/2$, it then follows that α_n goes to zero when n goes to infinity. Moreover, Jensen's inequality implies

$$\beta_n \leq 2\delta(\alpha_n + \sigma).$$

As both $\delta = \delta_n$ and α_n tend to zero when n goes to infinity, we have shown that

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n - u_n\|_n^2] = 0. \quad (21)$$

Remark that for any non negative random variable X ,

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(X \geq t) dt \leq n^{-1/4} + \int_0^{+\infty} \mathbb{P}(X \geq t) \mathbb{1}_{\{t \geq n^{-1/4}\}} dt.$$

From equation (24) in the proof of Lemma 3, we know that for any $t > 0$,

$$\mathbb{P}(|\|\hat{u}_n - u_n\|_n^2 - \|\hat{u}_n - u_n\|^2| \geq t) \leq \exp\left(2 \left\lceil \frac{64C^2}{t} \right\rceil \log n - \frac{t^2 n}{32C^2}\right).$$

Thus, setting

$$f_n(t) = \mathbb{1}_{[0, n^{-1/4}]}(t) + \exp\left(2 \left\lceil \frac{64C^2}{t} \right\rceil \log n - \frac{t^2 n}{32C^2}\right) \mathbb{1}_{\{t \geq n^{-1/4}\}},$$

we deduce that

$$\mathbb{E} [|\|\hat{u}_n - u_n\|_n^2 - \|\hat{u}_n - u_n\|^2|] \leq n^{-1/4} + \int_0^{+\infty} f_n(t) dt.$$

Then, it is readily seen that there exists an integer n_0 such that for all $n \geq n_0$ and for all $t \geq 0$, one has $f_n(t) \leq f_2(t)$. Since for all $t > 0$ fixed, $f_n(t)$ goes to 0 when n tends to infinity, it remains to invoke Lebesgue's dominated convergence Theorem to conclude

$$\mathbb{E} [\|\hat{u}_n - u_n\|_n^2] - \mathbb{E} [\|\hat{u}_n - u_n\|^2] \rightarrow 0.$$

Combining the latter with equation (21), we have obtained

$$\lim_{n \rightarrow \infty} \mathbb{E} [\|\hat{u}_n - u_n\|]^2 = 0,$$

which ends the proof of Proposition 2.

5.3 Proof of Proposition 3

Consider the translated cone

$$r + \mathcal{C}^+ = \{r + u, u \in \mathcal{C}^+\}.$$

Figure 4 provides a very simple interpretation of the algorithm: namely, it illustrates that the sequences of functions $u^{(k)}$ and $r - b^{(k)}$ might be seen as alternate projections onto the cones \mathcal{C}^+ and $r + \mathcal{C}^+$. In what follows, we justify this illuminating geometric interpretation in a rigorous way, and we explain its key role in the proof of the convergence as k goes to infinity.

By definition, we have $u^{(1)} = P_{\mathcal{C}^+}(r)$ where $P_{\mathcal{C}^+}$ denotes the metric projection onto \mathcal{C}^+ . Classical properties of projections ensure that

$$P_{r+\mathcal{C}^+}(u^{(1)}) = r + P_{\mathcal{C}^+}(u^{(1)} - r) = r - P_{\mathcal{C}^-}(r - u^{(1)}).$$

Coming back to the definition of $b^{(1)} = P_{\mathcal{C}^-}(r - u^{(1)})$, we are led to

$$r - b^{(1)} = P_{r+\mathcal{C}^+}(u^{(1)}).$$

In the same manner, since $u^{(2)} = P_{\mathcal{C}^+}(r - b^{(1)})$, we get

$$r - b^{(2)} = r - P_{\mathcal{C}^-}(r - u^{(2)}) = r + P_{\mathcal{C}^+}(r - u^{(2)}) = P_{r+\mathcal{C}^+}(u^{(2)}).$$

More generally, denoting $b^{(0)} = 0$, this yields for all $k \geq 1$ (see also figure 4)

$$u^{(k)} = P_{\mathcal{C}^+}(r - b^{(k-1)}) \quad \text{and} \quad r - b^{(k)} = P_{r+\mathcal{C}^+}(u^{(k)}).$$

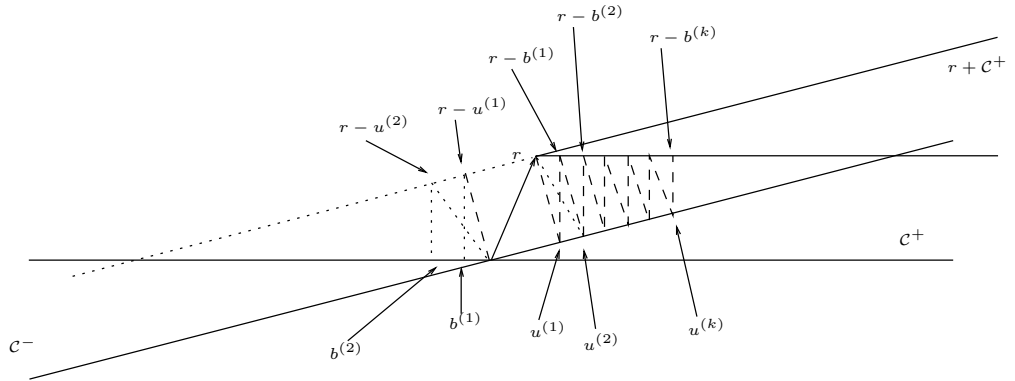


Figure 4: Interpretation of I.I.R. as a Von Neumann's algorithm.

It remains to invoke Theorem 4.8 in Bauschke and Borwein [4] to conclude that

$$(r - b^{(k)}) - u^{(k)} = r - r^{(k)} \xrightarrow[k \rightarrow \infty]{} 0,$$

which ends the proof of Proposition 3.

6 Technical results

6.1 Concentration inequalities

Throughout the previous proofs, we repeatedly needed to pass from the empirical norm $\|\cdot\|_n$ to the $L_2(\mu)$ norm $\|\cdot\|$. This was made possible thanks to several exponential inequalities that we justify in this section.

Specifically, let g and h denote two mappings from $I = [0, 1]$ to $[-C, C]$, and consider the random variable

$$\Delta_n(g - h) = \frac{1}{n} \sum_{i=1}^n \{(g(X_i) - h(X_i))^2 - \mathbb{E}[(g(X) - h(X))^2]\} = \|g - h\|_n^2 - \|g - h\|^2.$$

In what follows, we focus on the concentration of $\Delta_n(g - h)$ around zero. The first result is a straightforward application of Hoeffding's inequality.

Lemma 1 *For any couple of mappings g and h from $[0, 1]$ to $[-C, C]$, there exist positive real numbers α , β , c_1 and c_2 , depending only on C , and such that*

$$\mathbb{P}(|\Delta_n(g - h)| > n^{-\alpha}) \leq c_1 \exp(-c_2 n^\beta).$$

Proof. Since $|g(X_i) - h(X_i)| \leq 2C$, Hoeffding's inequality gives for all $t > 0$

$$\mathbb{P}(|\Delta_n(g - h)| > t) \leq 2 \exp\left(-\frac{t^2 n}{8C^2}\right) \quad (22)$$

Taking $t = n^{-\alpha}$ with $\alpha \in (0, 1/2)$, we deduce

$$\mathbb{P}(|\Delta_n(h)| > n^{-\alpha}) \leq 2 \exp\left(-\frac{n^{1-2\alpha}}{8C^2}\right)$$

and the result is proved with $c_1 = 2$, $c_2 = 1/(8C^2)$ and $\beta = 1 - 2\alpha > 0$. \square

The next lemma goes one step further, by considering, for fixed g , the tail distribution of

$$\sup_{h \in \mathcal{C}_{[0,1]}^+} |\Delta_n(g - h)|.$$

For obvious reasons, this type of result is sometimes called a maximal inequality. The proof shares elements with the one of Theorem 3.1 of van de Geer and Wegkamp [13].

Lemma 2 *Let g be a function from $[0, 1]$ to $[-C, C]$ and let $\mathcal{C}_{[0,1]}^+$ denote the set of non-decreasing functions from $[0, 1]$ to $[-C, C]$. There exist positive real numbers α' , β' , c'_1 and c'_2 depending only on C and such that*

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}_{[0,1]}^+} |\Delta_n(g - h)| > n^{-\alpha'}\right) \leq c'_1 \exp(-c'_2 n^{\beta'}).$$

Proof. The first step consists in showing that the mapping $h \mapsto \Delta_n(g - h)$ is Lipschitz. For any pair of functions h and \tilde{h} , we have

$$\begin{aligned} \Delta_n(g - h) - \Delta_n(g - \tilde{h}) &= \frac{1}{n} \sum_{i=1}^n \left\{ 2g(X_i) - h(X_i) - \tilde{h}(X_i) \right\} \left(\tilde{h}(X_i) - h(X_i) \right) \\ &\quad - \mathbb{E} \left[\left\{ 2g(X) - h(X) - \tilde{h}(X) \right\} \left(\tilde{h}(X) - h(X) \right) \right]. \end{aligned}$$

Since h and \tilde{h} take values in $[-C, C]$, we get

$$|\Delta_n(g - h) - \Delta_n(g - \tilde{h})| \leq 4C \times \left\{ \frac{1}{n} \sum_{i=1}^n |h(X_i) - \tilde{h}(X_i)| + \mathbb{E} [|h(X) - \tilde{h}(X)|] \right\}$$

and according to Jensen's inequality,

$$|\Delta_n(g - h) - \Delta_n(g - \tilde{h})| \leq 4C \times \left\{ \|h - \tilde{h}\|_n + \|h - \tilde{h}\| \right\}.$$

Now, since $\|h - \tilde{h}\| = \mathbb{E} [\|h - \tilde{h}\|_n]$, if the inequality $\|h - \tilde{h}\|_n \leq \delta$ is satisfied, we also have $\|h - \tilde{h}\| \leq \delta$. Thus,

$$\forall \delta > 0, \quad \|h - \tilde{h}\|_n \leq \delta \Rightarrow |\Delta_n(g - h) - \Delta_n(g - \tilde{h})| \leq 8C\delta$$

and the mapping $h \mapsto \Delta_n(g - h)$ is Lipschitz for the empirical norm $\|\cdot\|_n$.

Next, let us consider a δ -covering $\mathcal{E}^* = \{e_j^*, j = 1, \dots, M\}$ of $\mathcal{C}_{[0,1]}^+$ for the empirical norm $\|\cdot\|_n$. We stress that this set \mathcal{E}^* is random since it depends on the points X_i , but its cardinality M may be chosen deterministic and upper-bounded as follows (see Lemma 4): denoting $N = \lceil \frac{2C}{\delta} \rceil$, where $\lceil \cdot \rceil$ stands for the ceiling function, we have

$$M = \binom{n + N}{N} \leq n^N, \tag{23}$$

where the last inequality is satisfied for any integer $n \geq 2$ as soon as $N \geq 3$.

Then, for any h in $\mathcal{C}_{[0,1]}^+$, there exists e^* in \mathcal{E}^* such that $\|h - e^*\|_n \leq \delta$. From the previous Lipschitz property, we know that

$$|\Delta_n(g - h) - \Delta_n(g - e^*)| \leq 8C\delta.$$

Letting $t > 0$ and $\delta = t/(16C)$, our objective is to upper bound

$$\mathbb{P} \left(\sup_{h \in \mathcal{C}_{[0,1]}^+} |\Delta_n(g - h)| > t \right).$$

In this aim, for any h in $\mathcal{C}_{[0,1]}^+$ and any e^* in \mathcal{E}^* , we start with the decomposition

$$|\Delta_n(g - h)| \leq |\Delta_n(g - h) - \Delta_n(g - e^*)| + |\Delta_n(g - e^*)|.$$

For any h such that $|\Delta_n(g - h)| > t$, since there exists e^* in \mathcal{E}^* such that

$$|\Delta_n(g - h) - \Delta_n(g - e^*)| \leq t/2,$$

we necessarily have $|\Delta_n(g - e^*)| > t/2$, and consequently

$$\mathbb{P}(|\Delta_n(g - h)| > t) \leq \mathbb{P}\left(\max_{j=1 \dots M} |\Delta_n(g - e_j^*)| > t/2\right).$$

In other words,

$$\begin{aligned} \mathbb{P}\left(\sup_{h \in \mathcal{C}_{[0,1]}^+} |\Delta_n(g - h)| > t\right) &\leq \mathbb{P}\left(\max_{j=1 \dots M} |\Delta_n(g - e_j^*)| > t/2\right) \\ &\leq \mathbb{P}\left(\bigcup_{j=1}^M |\Delta_n(g - e_j^*)| > t/2\right) \\ &\leq \sum_{j=1}^M \mathbb{P}(|\Delta_n(g - e_j^*)| > t/2). \end{aligned}$$

According to (22) and to the fact that

$$M \leq n^N = n^{\lceil \frac{2C}{\delta} \rceil},$$

fixing $\delta = t/(16C)$ leads to

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}_{[0,1]}^+} |\Delta_n(g - h)| > t\right) \leq 2M \exp\left(-\frac{t^2 n}{8C^2}\right) \leq 2 \exp\left(\left\lceil \frac{32C^2}{t} \right\rceil \log n - \frac{t^2 n}{32C^2}\right).$$

Finally, for any $\alpha' \in (0, 1/3)$, there exists $c'_2 = c'_2(\alpha')$ such that for any integer n ,

$$\left\lceil \frac{32C^2}{n^{-\alpha'}} \right\rceil \log n - \frac{n^{-2\alpha'} n}{32C^2} \leq -c'_2 n^{1-2\alpha'},$$

hence the desired result with $t = n^{-\alpha'}$ and $\beta' = 1 - 2\alpha'$. \square

The last concentration inequality is a generalization of the previous one: this time, neither g nor h are assumed fixed.

Lemma 3 *Let us denote $\mathcal{C}_{[0,1]}^+$ the set of non decreasing mappings from $[0, 1]$ to $[-C, C]$. There exist positive real numbers α'' , β'' , c_1'' and c_2'' , depending only on C , and such that*

$$\mathbb{P} \left(\sup_{h_1 \in \mathcal{C}_{[0,1]}^+, h_2 \in \mathcal{C}_{[0,1]}^+} |\Delta_n(h_1 - h_2)| > n^{-\alpha''} \right) \leq c_1'' \exp \left(-c_2'' n^{\beta''} \right).$$

Proof. With the same notations as before, just note that for any mapping $h_1 \in \mathcal{C}_{[0,1]}^+$ (respectively h_2), there exists h_1^* (respectively h_2^*) in the δ -covering \mathcal{E}^* of $\mathcal{C}_{[0,1]}^+$, such that

$$\|h_1 - h_1^*\|_n \leq \delta \quad \text{and} \quad \|h_2 - h_2^*\|_n \leq \delta.$$

Following the same line as in the proof of the previous lemma, we have, for any mapping g with values in $[-C, C]$, that

$$|\Delta_n(g - h_1) - \Delta_n(g - h_1^*)| \leq 8C\delta \quad \text{and} \quad |\Delta_n(g - h_2) - \Delta_n(g - h_2^*)| \leq 8C\delta.$$

In particular

$$|\Delta_n(h_2 - h_1) - \Delta_n(h_2 - h_1^*)| \leq 8C\delta \quad \text{and} \quad |\Delta_n(h_1^* - h_2) - \Delta_n(h_1^* - h_2^*)| \leq 8C\delta.$$

Moreover,

$$|\Delta_n(h_1 - h_2)| \leq |\Delta_n(h_2 - h_1) - \Delta_n(h_2 - h_1^*)| + |\Delta_n(h_2 - h_1^*)|.$$

Set $\delta = t/(32C)$, then

$$|\Delta_n(h_1 - h_2)| > t \Rightarrow |\Delta_n(h_2 - h_1^*)| > 3t/4.$$

In the same manner,

$$|\Delta_n(h_2 - h_1^*)| \leq |\Delta_n(h_1^* - h_2) - \Delta_n(h_1^* - h_2^*)| + |\Delta_n(h_1^* - h_2^*)|,$$

and

$$|\Delta_n(h_2 - h_1^*)| > 3t/4 \Rightarrow |\Delta_n(h_1^* - h_2^*)| > t/2.$$

Hence, for any h_1 and h_2 in $\mathcal{C}_{[0,1]}^+$,

$$\mathbb{P} (|\Delta_n(h_1 - h_2)| > t) \leq \mathbb{P} \left(\max_{h_1^*, h_2^* \in \mathcal{E}^*} |\Delta_n(h_1^* - h_2^*)| > t/2 \right).$$

As a consequence, the choice $\delta = t/(32C)$ gives

$$\begin{aligned} \mathbb{P} \left(\sup_{h_1 \in \mathcal{C}_{[0,1]}^+, h_2 \in \mathcal{C}_{[0,1]}^+} |\Delta_n(h_1 - h_2)| > t \right) &\leq \mathbb{P} \left(\max_{h_1^*, h_2^* \in \mathcal{E}^*} |\Delta_n(h_1^* - h_2^*)| > t/2 \right) \\ &\leq \sum_{1 \leq j_1 \neq j_2 \leq M} \mathbb{P} (|\Delta_n(e_{j_1}^* - e_{j_2}^*)| > t/2) \\ &\leq M^2 \exp \left(-\frac{t^2 n}{32C^2} \right). \end{aligned}$$

According to (23), we are led to

$$\mathbb{P} \left(\sup_{h_1 \in \mathcal{C}_{[0,1]}^+, h_2 \in \mathcal{C}_{[0,1]}^+} |\Delta_n(h_1 - h_2)| > t \right) \leq \exp \left(2 \left\lceil \frac{64C^2}{t} \right\rceil \log n - \frac{t^2 n}{32C^2} \right). \quad (24)$$

For any $\alpha'' \in (0, 1/3)$, there exists a real number $c_2'' = c_2''(\alpha'')$ such that for any integer n

$$2 \left\lceil \frac{64C^2}{n^{-\alpha''}} \right\rceil \log n - \frac{n^{-2\alpha''} n}{32C^2} \leq -c_2'' n^{1-2\alpha''},$$

hence the desired result with $t = n^{-\alpha''}$ and $\beta'' = 1 - 2\alpha''$. \square

We conclude this section with the proof of inequality (23). It borrows elements from Lemma 3.2 in van de Geer [12].

Lemma 4 *Denote $\mathcal{C}_{[0,1]}^+$ the set of non-decreasing mappings from $[0, 1]$ to $[-C, C]$, and $\|\cdot\|_n$ the empirical norm with respect to the sample (X_1, \dots, X_n) . For any $\delta > 0$, there exists a δ -covering of $(\mathcal{C}_{[0,1]}^+, \|\cdot\|_n)$ with cardinality less than $M = \binom{n+N}{N}$, where $N = \lceil \frac{2C}{\delta} \rceil$, and $\lceil \cdot \rceil$ stands for the ceiling function.*

Proof. Let us rewrite $X_{(1)} \leq \dots \leq X_{(n)}$ the reordering of the sample (X_1, \dots, X_n) in increasing order. Recall that the empiric norm is defined for any pair of functions g and h in $\mathcal{C}_{[0,1]}^+$ by

$$\|g - h\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (g(X_{(i)}) - h(X_{(i)}))^2},$$

Hence, if $|g(X_{(i)}) - h(X_{(i)})| \leq \delta$ for all indices $i = 1, \dots, n$, we also have $\|g - h\|_n \leq \delta$.

For the sake of simplicity, let us assume that $N_0 = C/\delta$ is an integer and let us consider the following partition of the interval $[-C, C]$

$$\mathcal{S} = \{-C = -N_0\delta < -(N_0 - 1)\delta < \dots < -\delta < 0 < \delta < \dots < (N_0 - 1)\delta < N_0\delta = C\}.$$

Let us denote $\mathcal{I}_{[0,1]}^+$ the set of non-decreasing functions defined on $[0, 1]$, with values in \mathcal{S} and piecewise constant on the intervals $(X_{(i)}, X_{(i+1)})$. We also suppose that they are constant on the intervals $[0, X_{(1)}]$ and $[X_{(n)}, 1]$, with respective values the ones of $X_{(1)}$ and $X_{(n)}$.

Firstly, it is readily seen that any function g in $\mathcal{C}_{[0,1]}^+$ may be approximated at a distance less than or equal to δ with respect to the empirical norm $\|\cdot\|_n$ by a function in $\mathcal{I}_{[0,1]}^+$. For this, it indeed suffices to pick at each point $X_{(i)}$ the nearest value of $g(X_{(i)})$ in the

partition \mathcal{S} . Secondly, it is well-known in discrete mathematics (see for example Lovász *et al.* [22], Theorem 3.4.2) that

$$|\mathcal{I}_{[0,1]}^+| = \binom{n+N}{N}.$$

□

6.2 Proof of Lemma 5

Consider the subset $\mathcal{C}_{n,C}^+$ of \mathcal{C}_n^+ consisting in all vectors whose absolute values of the components are bounded by a real number C . Consider $N \in \mathbb{N}$ such that $N \leq n$. For each $j = 0, \dots, N-1$, let us introduce the vector $h_j^+ = (h_j^+[1], \dots, h_j^+[n])'$ of \mathbb{R}^n as follows

$$h_j^+[i] = \begin{cases} 0 & \text{if } i \leq \lfloor \frac{jn}{N} \rfloor \\ 1 & \text{otherwise} \end{cases}$$

and define

$$H_+ = \text{Vect}(h_0^+, \dots, h_{N-1}^+).$$

Finally, set $\delta = 2\sqrt{2}C/\sqrt{N} \geq 2\sqrt{2}C/\sqrt{n}$.

Lemma 5 *With the previous notations, we have for all f in $\mathcal{C}_{n,C}^+$*

$$\inf_{h \in H_+} \|f - h\|_n \leq \delta.$$

Proof. We denote $f = (f[1], \dots, f[n])'$, with

$$-C \leq f[1] \leq \dots \leq f[n] \leq C.$$

Set $\alpha_N = f[n]$ and, for $j = 0, \dots, N-1$,

$$\alpha_j = \min_{i: h_j^+[i]=1} f[i]$$

We define also the vectors f_- and f_+ of H_+ as follows

$$f_- = \alpha_0 h_0^+ + \sum_{j=1}^{N-1} (\alpha_j - \alpha_{j-1}) h_j^+$$

and

$$f_+ = \alpha_1 h_0^+ + \sum_{j=1}^{N-1} (\alpha_{j+1} - \alpha_j) h_j^+.$$

Then we note that $f_- \leq f \leq f_+$, so that

$$\|f - f_-\|_n^2 \leq \|f_+ - f_-\|_n^2$$

with

$$f_+ - f_- = \sum_{j=1}^{N-1} (\alpha_j - \alpha_{j-1})(h_{j-1}^+ - h_j^+) + (\alpha_N - \alpha_{N-1})h_{N-1}^+. \quad (25)$$

Remark that, for all $j = 1, \dots, N-1$,

$$\|h_{j-1}^+ - h_j^+\|_n^2 \leq \frac{1}{n} \left(\lfloor \frac{jn}{N} \rfloor - \lfloor \frac{(j-1)n}{N} \rfloor \right) \leq \frac{1}{n} \left(\frac{n}{N} + 1 \right) \leq \frac{2}{N},$$

and $\|h_{N-1}^+\|_n^2 \leq 2/N$ as well. Thus, taking into account that the decomposition (25) is orthogonal, we get

$$\|f_+ - f_-\|_n^2 \leq \frac{2}{N} \sum_{j=1}^N (\alpha_j - \alpha_{j-1})^2 = \frac{8C^2}{N} \sum_{j=1}^N \left(\frac{\alpha_j - \alpha_{j-1}}{2C} \right)^2.$$

Since $0 \leq (\alpha_j - \alpha_{j-1})/(2C) \leq 1$ and $0 \leq (\alpha_N - \alpha_1)/2C \leq 1$, we have

$$\|f_+ - f_-\|_n^2 \leq \frac{8C^2}{N} \sum_{j=1}^N \frac{\alpha_j - \alpha_{j-1}}{2C} \leq \frac{8C^2}{N}.$$

Considering that $\delta^2 = 8C^2/N$, we finally get the desired result, that is

$$\inf_{h \in H_+} \|f - h\|_n^2 \leq \delta^2.$$

□

For the subset $\mathcal{C}_{n,C}^-$ of \mathcal{C}_n^- , we proceed in the same way. We conclude that there exists a vector space H_- with dimension $N = 8C^2/\delta^2$ such that, for all f in $\mathcal{C}_{n,C}^-$,

$$\inf_{h \in H_-} \|f - h\|_n \leq \delta.$$

Acknowledgments. We wish to thank Dragi Anevski and Enno Mammen to have made us aware of reference [1]. Arnaud Guyader is greatly indebted to Bernard Delyon for fruitful discussions on Von Neumann's algorithm.

References

- [1] D. Anevski and P. Soulier (2011). Monotone spectral density estimation. *The Annals of Statistics*, **39**(1), 418-438.
- [2] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641-647.

- [3] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk (1972). *Statistical inference under order restrictions: Theory and application of isotonic regression*. John Wiley & Sons.
- [4] H.H. Bauschke and J.M. Borwein (1994). Dykstra's alternating projection algorithm for two sets. *Journal of Approximation Theory*, **79**(3), 418-443.
- [5] M.J. Best and N. Chakravarti (1990). Active set algorithms for isotonic regression; An unifying framework. *Mathematical Programming*, **47**(1), 425-439.
- [6] H.D. Brunk (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 607-616.
- [7] H.D. Brunk (1970). Estimation of isotonic regression. *Cambridge University Press*, 177-195.
- [8] A. Buja, T.J. Hastie, and R.J. Tibshirani (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**(2), 453-510.
- [9] C. Durot (2007). On the Lp-error of monotonicity constrained estimators. *The Annals of Statistics*, **35**(3), 1080-1104.
- [10] R.L. Dykstra (1981). An isotonic regression algorithm. *Journal of Statistical Planning and Inference*, **5**(4), 355-363.
- [11] J.H. Friedman and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 817-823.
- [12] S. van de Geer (1987). A new approach to least-squares estimation, with applications. *The Annals of Statistics*, **15**(2), 587-602.
- [13] S. van de Geer and M. Wegkamp (1996). Consistency for the least squares estimator in nonparametric regression. *The Annals of Statistics*, **24**(6), 2513-2523.
- [14] A. Guyader, N. Jégou, A.B. Németh, and S.N. Németh (2012). A Geometrical Approach to Iterative Isotone Regression. <http://arxiv.org/abs/1211.3930>
- [15] D.L. Hanson, G. Pledger, and F.T. Wright (1973). On consistency in monotonic regression. *The Annals of Statistics*, **1**(3), 401-421.
- [16] W. Härdle and P. Hall (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**(1), 43-57.
- [17] T.J. Hastie and R.J. Tibshirani (1990). *Generalized additive models*. Chapman & Hall/CRC.
- [18] J. Horowitz, J. Klemelä, and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli*, **12**(2), 271-298.

- [19] N.W. Hengartner and S. Sperlich (1999). Rate optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, **95**(2), 246-272.
- [20] W. Kim, O.B. Linton, and N.W. Hengartner (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**(2), 278-297.
- [21] C.I.C. Lee (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics*, **11**(2), 467-477.
- [22] L. Lovász, J. Pelikán, and K. Vesztergombi (2003). *Discrete Mathematics: Elementary and Beyond*. Springer-Verlag, New York.
- [23] E. Mammen, O. Linton, and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, **27**(5), 1443-1490.
- [24] E. Mammen and K. Yu (2007). Additive isotone regression. *Asymptotics: Particles, Processes and Inverse Problems, IMS Lecture Notes-Monograph Series*, **55**, 179-195.
- [25] M. Meyer and M. Woodroffe (2000). On the Degrees of Freedom in Shape-Restricted Regression. *The Annals of Statistics*, **28**(4), 1083-1104.
- [26] J.D. Opsomer and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**(1), 186-211.
- [27] J.D. Opsomer (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**(2), 166-179.
- [28] T. Robertson, F.T. Wright, and R.L. Dykstra (1988). *Order Restricted Statistical Inference*. Wiley, New York.
- [29] C.J. Stone (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2), 689-705.